KMI, The Open University at NTCIR-9 CrossLink: Cross-Lingual Link Discovery in Wikipedia using explicit semantic analysis

Conference or Workshop Item

KMI, The Open University at NTCIR-9 CrossLink: Cross-Lingual Link Discovery in Wikipedia Using Explicit Semantic Analysis

Petr Knoth
KMI, The Open University
Walton Hall, Milton Keynes
United Kingdom
p.knoth@open.ac.uk

Lukas Zilka
KMI, The Open University
Walton Hall, Milton Keynes
United Kingdom
l.zilka@open.ac.uk

Zdenek Zdrahal
KMI, The Open University
Walton Hall, Milton Keynes
United Kingdom
z.zdrahal@open.ac.uk

ABSTRACT

This paper describes the methods used in the submission of Knowledge Media institute (KMI), The Open University to the NTCIR-9 Cross-Lingual Link Discovery (CLLD) task entitled CrossLink. KMI submitted four runs for link discovery from English to Chinese; however, the developed methods, which utilise Explicit Semantic Analysis (ESA), are applicable also to other language combinations. Three of the runs are based on exploiting the existing cross-lingual mapping between different versions of Wikipedia articles. In the fourth run, we assume information about the mapping is not available. Our methods achieved encouraging results and we describe in detail how their performance can be further improved. Finally, we discuss two important issues in link discovery: the evaluation methodology and the applicability of the developed methods across different textual collections.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*; I.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*linguistic processing*

General Terms

Algorithms, Experimentation, Languages

Keywords

Cross-lingual Link Discovery, Link Discovery, Semantic Similarity, Explicit Semantic Analysis, NTCIR, Wikipedia

1. INTRODUCTION

Cross-referencing documents is an essential part of organising textual information. However, keeping links in large, quickly growing document collections up-to-date, is problematic due to the number of possible connections. In multilingual document collections, interlinking semantically related information in a timely manner becomes even more challenging. Suitable software tools that could facilitate the link discovery process by automatically analysing the multilingual content are currently not not available.

In this paper, we present Cross-Lingual Link Discovery (CLLD) methods that can be used to suggest a set of cross-lingual links from an English Wikipedia article to articles in another language. Our experiments were carried out on English to Chinese, but the methods are applicable also to other language combinations.

2. RELATED WORK

Current approaches to link detection can be divided into three groups:

(1) *link-based* approaches discover new links by exploiting the existing link graph [Jenkinson et al.,2008; Lu et al.,2008].

(2) *semi-structured* approaches try to discover new links using semi-structured information, such as the anchor texts or document titles [Geva,2007; Dopichaj et al.,2008; Granitzer et al.,2008; Milne and Witten,2008; Mihalcea and Csomai,2007].

(3) *purely content-based* approaches use as an input plain text only. They typically discover related resources by calculating semantic similarity based on document vectors [Allan,1997; Green,1998; Zeng and Bloniarz,2004; Zhang and Kamps,2008; He,2008; Knoth et al.,2010]. Some approaches, such as [Itakura and Clarke,2008; Lu et al.,2008; Knoth et al.,2011a], combine multiple approaches.

A major disadvantage of the link-based and semi-structured approaches is the difficulty associated with porting them across different types of document collections. The two well-known solutions to monolingual link detection, the Geva's and Itakura's algorithms [Trotman et al.,2009], fit in these two categories. While these algorithms have been demonstrated to be effective on a specific Wikipedia set, their performance has significantly decreased when they were applied to a slightly different task of interlinking two encyclopedia collections. Purely content-based methods have been mostly found to produce slightly worse results than the two previous classes of methods; however, their advantage is that their performance should in theory remain stable across different document collections. As a result, they can always be used as part of any link discovery system and can even be combined with domain specific methods that make use of the link graph or semi-structured information. In practice, domain-specific link discovery systems can achieve high

precision and recall. For example, *Wikify!* [Mihalcea and Csomai,2007] and the link detector presented by [Milne and Witten,2008] can be used to identify suitable anchors in text and enrich it with links to Wikipedia by combining multiple approaches with domain knowledge.

NTCIR-9 CrossLink is the first evaluation forum to stimulate the development and compare the performance of multilingual link discovery systems for Wikipedia. The methods submitted by different teams often build on top of successful monolingual systems and solutions, exploiting mainly semi-structured and link-based information, adapted for the multilingual task. The most common approaches to the issue of multilingualism are (a) using the manually defined mapping between equivalent Wikipedia articles or (b) using machine translation. In one of our runs, we have also explored the possibility of applying cross-language semantic similarity measures.

In this paper, we describe four methods for CLLD submitted by KMI, all of which utilize Explicit Semantic Analysis (ESA). Measuring semantic similarity using ESA has been previously found to produce better results than calculating it directly on document vectors using cosine and other similarity measures and it has also been found to outperform the results that can be obtained by measuring similarity on vectors produced by Latent Semantic Analysis (LSA) [Gabrilovich and Markovitch,2007]. Therefore, the application of ESA seems a plausible choice.

3. LINK DISCOVERY METHODS

The CLLD methods we have developed operate in three phases: *target discovery*, *anchor detection* and *link ranking*, as demonstrated in Figure 1. In the first phase, we take the orphan document (topic) in the original language and try to find other documents in the target language that could be considered suitable link targets, using semantic similarity as a criterion. In the second step, we take the list of candidate targets and try to detect for each of them a suitable anchor in the orphan document. In the third phase, we describe each anchor using a set of features and perform link ranking using a Support Vector Machine (SVM) classifier. The following subsection describes them in more detail.

3.1 Target discovery

In the target discovery phase we take as an input a new "orphan" document (i.e. a document that is not linked to other documents) written in the source language and we automatically generate a list of potential target documents. In this phase, the system works at the granularity of the whole documents.

We apply two different approaches to accomplish this task. The first approach is based on the application of ESA in combination with the existing link structure of Wikipedia, and we will call it *ESA2Links*. The second approach utilises the information about page titles on Wiki pages, and we will call it *Terminology*. Both approaches can be combined or used separately.

The *ESA2Links* method works in two steps. In the first step, an ESA vector is calculated for each document in the document collection. This results in obtaining a weighted vector of Wikipedia concepts for each document in the source language. The cardinality of the vector is given by the number of concepts (i.e. pages) in the source language version of Wikipedia (about 3.8 million for English). The same procedure is applied on the orphan document. Similarity between the resulting ESA vectors is then calculated and k most similar pages are identified. In our experiments we use $k = 1,000$.

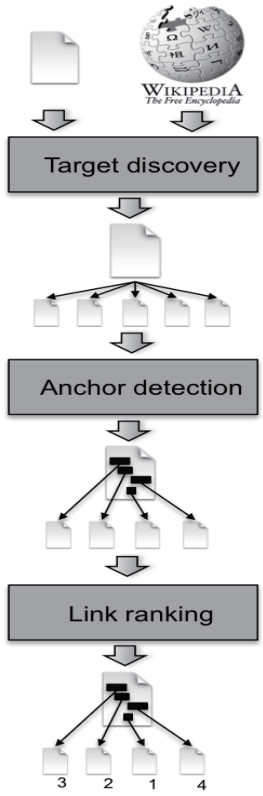

Figure 1: Cross-Lingual Link Discovery process

In the second step, the k most similar documents to the orphan document are taken as a seed and are used to discover documents that are suitable link targets. In our previous paper [Knoth et al.,2011b], we have described and evaluated four alternative approaches to target discovery. The approach producing the best results have been used. This approach requires access to the link structure in the document collection (please see [Knoth et al.,2011b] for alternative approaches that do not have this requirement). After generating the seed documents, the methods extracts all links in the form $[anchor, pageID]$ present in those seed documents, where $pageID$ is the Wiki identifier of the anchor destination. Using the cross-lingual mapping between Wikipedia pages, the $pageID$, describing a page in the source language, is mapped to an appropriate ID describing the same page in the target language. If the mapping is not explicitly specified in Wikipedia, the link is discarded. The resulting set of pairs represents the set of candidate targets.

The *Terminology* approach is much simpler than the previous one and can be considered the baseline approach. The method exploits the title information of Wiki articles and the cross-lingual mapping between Wikipedia articles. The method recommends as targets all pairs $[pageTitle, pageID]$ in the whole Wikipedia for which there exists an explicit cross-lingual mapping between the source and the target language version of the page, i.e. the resulting set of targets

will be always the same regardless of the orphan document. It is up to the next phase to filter down the list of targets to those that are suitable.

3.2 Anchor detection

In the anchor detection phase, we take as an input the set of targets and try to detect suitable anchors for them in the orphan document. The procedure is quite simple: We iterate through the set of target documents and we try to find a suitable anchor text in the orphan document given the target document title. If no anchor is discovered, the link is discarded.

The simplicity of this phase is very much given by the fact that the methods are tailored for Wikipedia. In Wikipedia, each page is characterised by a title. In addition, the anchor texts in Wikipedia are typically identical to the name of the title of the page which describes a given concept or are variations of the title which can easily be extracted from the collection. This is not the case in general (non-Wiki style) text collections where this step is significantly more challenging given the variability of link types [Knoth and Zdrahal,2011].

3.3 Link ranking

In the link ranking phase, we take the list of links in the form $[anchor, targetID]$, where *anchor* represents the specific text in the orphan document and the $targetID$ is the Wiki *page ID* of the target page in the target language, and we rank the links according to their importance defined as the confidence of the ranking system.

The approach we are using to generate our runs is based on machine learning. Each link is first described and modelled by a set of features (occurrence, generality and link frequency are inspired by [Milne and Witten,2008]). The features are represented as a vector assuming their mutual independence.

- *ESA similarity* is a real number between 0 and 1, which expresses the similarity of text. Three different features were included:
  - Similarity of the link text to the target document text.
  - Similarity of the link text to the target document title.
  - Similarity of the input document text to the target document text.
- *Generality* is a measure expressing how general a given topic is. It is an integer number between 0 and 16 defined as the minimum depth at which the topic is located in Wikipedia's category tree.
- *Link frequency* is a measure expressing how many times a particular keyword occurs as a link (or more precisely as an anchor) in the whole document collection.
- *Occurrence of the link text in the input document* is a relative measure of the first, last and current occurrence of the link text in the input document, and the difference between its first and last occurrence.

When the features are encoded, we train a Support Vector Machine (SVM). In our experiments, the system was trained on the examples associated to the three topic documents provided by the NTCIR CrossLink organisers. Negative examples were acquired by running the *ESA2Links* and anchor detection method described above, and by filtering out the positive examples provided by the organisers. In the testing phase the SVM classifier is used to decide whether a link should be included. Given the low number of training examples, we expect the SVM to have a relatively low recall, but high precision. The confidence value of the SVM, which characterises the distance from the decision hyperplane, is used to select the best candidates. The candidates are then ranked according to their semantic similarity to the orphan document.

3.4 Cross-lingual discovery

KMI has submitted four runs out of which three use the explicit information about cross-lingual mapping between Wiki pages. This makes the methods more difficult to reuse in other contexts. As a result, we have also tested in one of our runs a more challenging setting in which we utilise Cross-Lingual Explicit Semantic Analysis (CL-ESA) to discover an equivalent page in the target language (Chinese) for a page in the source language (English). The method is based on the mapping of the ESA conceptual space between the two languages. In our runs, we refer to this approach as *ESA discovery*.

The most semantically similar target language document to the orphan document is considered by the method as a suitable candidate. To identify such a document, cosine similarity is calculated between the ESA vector of the source document with the ESA vectors of other documents in the target document collection.

Each dimension in an ESA vector expresses the similarity of a document to the given language version of a Wikipedia concept/article. Therefore, the cardinality of the source document vector is different from the cardinality of the vectors representing the documents in the target language collection (see Figure 2). In order to calculate the similarity of two vectors, we map the dimensions that correspond to the same Wikipedia concepts in different language versions. In most cases, if a Wikipedia concept is mapped to another language version, there is a one-to-one correspondence between the articles in those two languages. However, there are cases when one page in the source language is mapped to more than one page in the target language and vice versa.[1] For the purpose of similarity calculation, we use 100 dimensions with the highest weight that are mappable from the source to the target language. We use the same number of dimensions in monolingual ESA in the target discovery phase.

3.5 KMI runs

KMI has submitted four runs for NTCIR CrossLink for English to Chinese. The methods are applicable to other language combinations, but we have tested on Chinese only.

Run 1: *KMI_SVM_ESA_TERMDB* - combines *ESA2Links* with *Terminology*

Run 2: *KMI_SVM_ESA* - applies *ESA2Links* for target discovery.

[1] These multiple mappings appear quite rarely, e.g. in 5,889 cases out of 550,134 for Spanish to English.

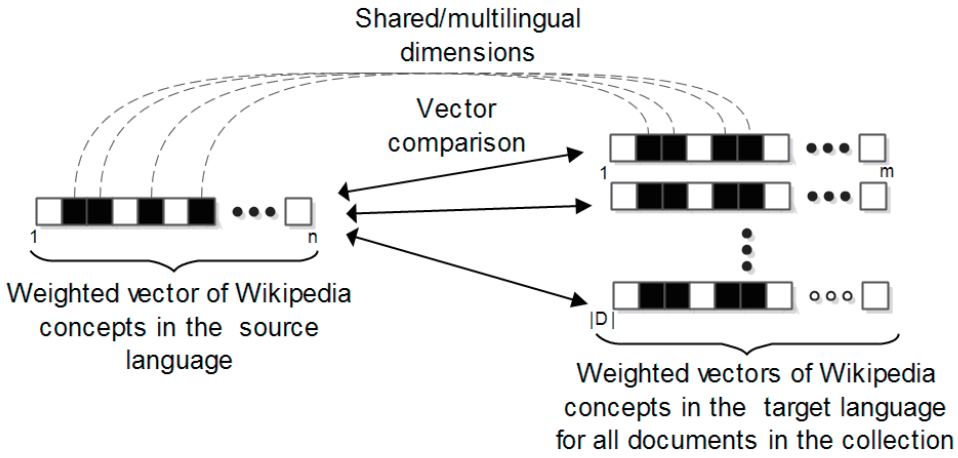

Figure 2: Calculating similarity of texts in different languages using CL-ESA

Run 3: *KMI_SVM_TERMDB* - uses *Terminology* only for target discovery.

Run 4: *KMI_ESA_SVM_ESAdiscovery* - uses *ESA2Links* for target discovery and ESA discovery for the cross-language step.

4. EXPERIMENTS

4.1 Evaluation methodology

All links and supporting information were cleared from the English articles used in the evaluation. The remaining link structure has been kept. The methods have been evaluated at different granularity levels anchor-to-file (A2F) and file-to-file (F2F). There were two evaluation modes:

- Wikipedia ground truth - the ground truth is derived automatically from the existing link structure of Wikipedia.
- Manual assessment - all anchors and targets are pooled and the evaluation is carried out by a human assessor.

Precision-at-N (P@N), *R-Prec*, and *Mean Average Precision (MAP)* were used as the main metrics to evaluate the performance of the CLLD methods. More information about the ground truth, the evaluation setup and a detailed description of the evaluation measures can be found in the overview paper [Tang et al.,2011].

4.2 Evaluation

All four KMI runs were submitted for English to Chinese. The F2F performance of the KMI methods with Wikipedia ground truth is shown in Figure 3. There is no A2F evaluation with Wikipedia ground truth as such evaluation would be difficult for a number of reasons: "An anchor can occur multiple times in a document in subtly different linguistic forms. It is unreasonable to score multiple identical links and also unreasonable not to score different linguistic variants. The best approach to measuring this imprecision is unclear and has been studied at the INEX Link Discovery Track where it changed from year to year" [Tang et al.,2011].

Figure 4 and Figure 5 show the performance of the presented methods when manual assessment has been used for both F2F and A2F granularity levels. The results for all experiments are summarised in Table 1.

4.3 Comparison of the runs performance

Overall, we can see that the *KMI_SVM_ESA_TERMDB* method achieved the best results in terms of MAP and R-Prec in all evaluations. Very similar results have been achieved by the *KMI_SVM_ESA* method showing that the use of the terminology dictionary in the target discovery step helps only moderately. The *KMI_SVM_TERMDB* method produced in most cases substantially worse results than the two methods that used the *ESA2Links* approach. This shows that combining semantic similarity with the information about existing Wikipedia links provides valuable information.

It is not surprising that the *KMI_ESA_SVM_ESADiscovery* method produced on this dataset worse results than the other methods as it is the only method that makes use of the explicit (manually created) cross-language mapping between different language versions of Wikipedia articles. On the other hand, this method is more generally applicable than the other methods.

4.4 Performance comparison with other teams

The KMI methods scored first in Precision-at-5 in the A2F manual assessment and third in terms of R-Prec. Our methods were also second in F2F manual assessment in terms of MAP and R-Prec and third in terms of Precision-at-5. Our system ranked third in F2F Wikipedia ground-truth evaluation in terms of all MAP, R-Prec and Precision-at-5.

4.5 Unique relevant links

The CrossLink organisers decided this year to also compare the systems based on the number of unique relevant links the individual systems have contributed [Tang et al.,2011]. The KMI team ranked in the comparison third for Wikipedia ground-truth (27 unique relevant links) and second for manual assesment (152 unique relevant links). However, we believe the results of this comparison should be interpreted

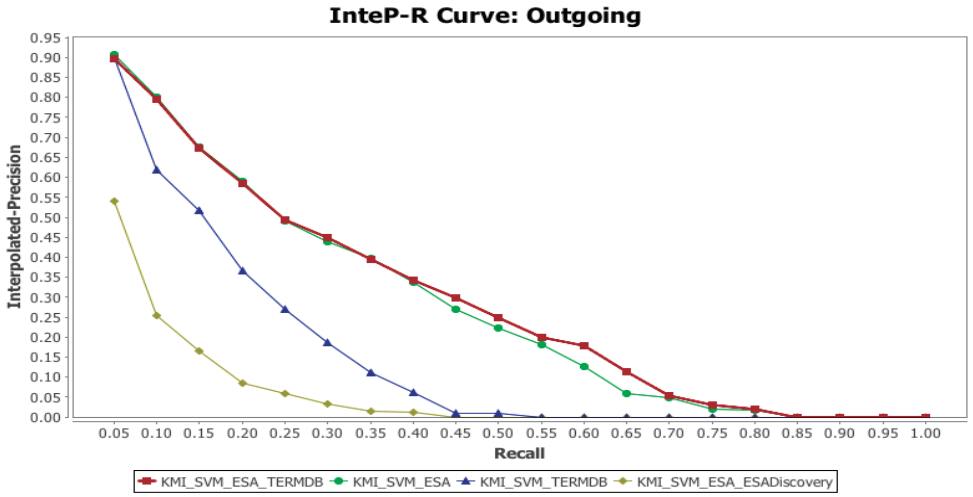

Figure 3: F2F performance of the KMI runs using Wikipedia ground truth.

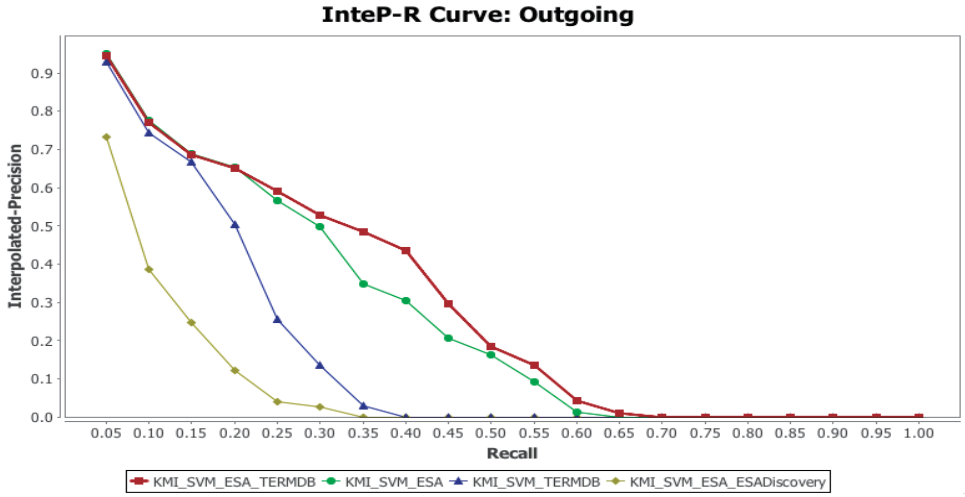

Figure 4: F2F performance of the KMI runs using manual assessment.

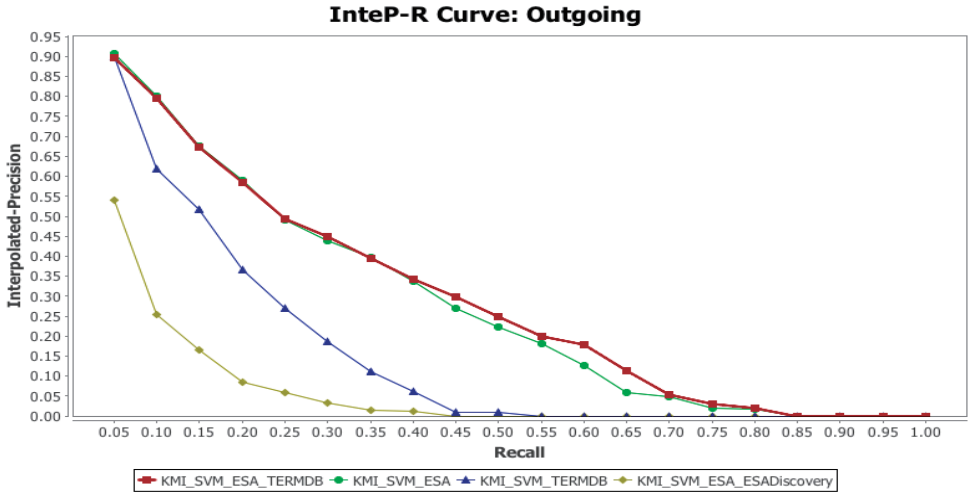

Figure 5: A2F performance of the KMI runs using manual assessment.

Run ID	MAP	R-Prec	P@5	P@10	P@20	P@30	P@50	P@250
F2F performance with Wikipedia ground truth								
KMI_SVM_ESA_TERMDB	**0.260**	**0.345**	0.712	**0.664**	0.530	0.491	**0.434**	**0.166**
KMI_SVM_ESA	0.251	0.338	**0.728**	**0.664**	**0.540**	**0.493**	0.430	0.153
KMI_SVM_TERMDB	0.127	0.211	0.624	0.552	0.454	0.383	0.302	0.078
KMI_ESA_SVM_ESADiscovery	0.059	0.148	0.264	0.240	0.186	0.165	0.138	0.044
F2F performance with manual assessment results								
KMI_SVM_ESA_TERMDB	**0.258**	**0.393**	0.720	**0.728**	**0.684**	0.648	0.604	**0.358**
KMI_SVM_ESA	0.231	0.344	0.728	0.720	0.678	**0.668**	**0.615**	0.306
KMI_SVM_TERMDB	0.133	0.192	**0.752**	0.692	0.636	0.613	0.561	0.178
KMI_ESA_SVM_ESADiscovery	0.054	0.132	0.464	0.388	0.348	0.321	0.283	0.119
A2F performance with manual assessment results								
KMI_SVM_ESA_TERMDB	**0.097**	**0.114**	**0.368**	**0.368**	**0.330**	0.303	0.269	**0.142**
KMI_SVM_ESA	0.080	0.092	0.360	0.364	**0.330**	0.299	0.260	0.113
KMI_SVM_TERMDB	0.070	0.075	0.376	**0.368**	0.324	**0.316**	**0.297**	0.096
KMI_ESA_SVM_ESADiscovery	0.014	0.035	0.088	0.108	0.110	0.108	0.090	0.045

Table 1: Performance of the KMI methods

very carefully because:

- This evaluation metric was not known to the participants prior to the submission and therefore the system parameters were not optimised to achieve high results in this evaluation.
- The results that are being compared are the number of unique links provided by the runs of different teams and therefore teams that have submitted less runs than the others are at a disadvantage.
- The comparison puts systems that have not generated all allowed 1,250 links per topic at a disadvantage. For example, a run producing high precision results can receive low score according to this measure in case it does not decide to generate all 1,250 links (a constant defined by the task organisers). Since this evaluation measure does not take into account the assigned rank to a particular link, a system that has generated, for example, 200 good links will recieve a lower score than a system that has generated first 1,000 links wrongly and the last 250 links are correct.
- It should be expected that the number of unique relevant links generated can differ significantly based on the selection and variation of parameters of different systems. Therefore in the future, such an evaluation should be carried out by taking into account the sensitivity of the systems to parameters, and the trade-off between unique relevant links and precision/recall characteristics.

4.6 How can the performance be improved?

There is a number of ways in which our methods could be improved and optimised for better performance. We see the main possibilities in:

Extending the set of training examples - the link ranking system (step 3) has been trained on a very limited number of examples. These examples included links relevant to only three topic documents provided by the organisers (i.e. Australia, Femme fatale and Martial arts). It is therefore reasonable to assume that just a moderately larger training data could increase the ranker performance.

Extending the methods to enable linking all articles - The three best performing methods we have presented rely on the existence of cross-lingual links in the Wikipedia collection. Our experiments show that for a large proportion of Chinese articles the mapping to English is missing. Therefore, our methods could be improved if this information was present in Wikipedia or by using methods that can detect different language versions for a Wikipedia article. Such a method was, for example, presented in [Sorg and Cimiano,2008].

Dimensionality of the ESA vector - to be able to run the methods quickly on our machines we decided to represent each document using only the best 100 ESA dimensions. The other dimensions of the vector were set to zero. While our experiments show that preserving only the best 100 dimensions strongly correlates (0.825 Spearman's rank correlation) with the results produced with 10,000 dimensions, preserving 2,500 dimensions would result in an almost perfect correlation of 0.98. We can assume that this could slightly improve the results of the three best performing methods and significantly improve the results of the method that makes use of cross-language discovery using ESA (Run 4).

The cross-language discovery step - we analysed the method relying on cross-language discovery of Wikipedia articles (Run 4), in particular the step in which the system takes an English article and tries to automatically discover the version of the same article in the target language. This task is difficult as the system has to select the correct article given the set of all articles in the target language. For Chinese, this amounts to 318,736 documents. Our results indicate that the correct document is selected as the first one in only 13% of cases, however, in 40% of cases it is among the top 10 documents and in 75% of cases among the top 100. We believe that this is mainly due to the fact that there is often a significant difference between the description of the same concept (i.e. the text on a Wiki page describing the same concept) across language versions.

Unique relevant links - the results of our system in terms of unique relevant links could be significantly improved by: (a) generating all (1,250) links allowed by the organisers per topic (in most cases our system generated about 200 links per topic). This can be achived by changing the k parameter of the system controlling the number of articles used as a

seed in the target discovery step.

5. DISCUSSION

5.1 Evaluation methodology

Choosing the right evaluation methodology is certainly one of the greatest challenges in link discovery. Suitable interlinked corpora that could be used for evaluation is lacking and creating it manually would require huge effort. It has been previously reported [Huang et al.,2008] that Wikipedia should not be seen as a reliable ground truth. When establishing the ground truth based on the link structure of different language versions, we can see that the correlation of their link structures is surprisingly low [Knoth et al.,2011a]. Therefore the automatic assessment results should be treated as informative only. At the same time, care should also be taken when interpreting the manual assessment results as the task of interlinking content can be considered highly subjective [Knoth et al.,2011a].

5.2 Applicability to a non-Wiki context

The aim in NTCIR-9 CrossLink was to develop a system that is performing well on the provided Wikipedia collection. However, technology for automatic document cross-referencing is also essential in many non-Wiki style document collections. Therefore, it is a question how easy or difficult it is to apply the developed methods in their context.

The results of the previous link discovery evaluation workshops (Link the Wiki Track: 2007-2010) show that methods relying on the existence of links or semi-structured information that is unique to Wiki-style collections (for example, the correspondence of anchors to article titles) is superior to the methods that are based on purely textual information. Therefore, it is not surprising to see that the majority of the submitted runs to NTCIR 2011: CrossLink were generated by systems exploiting the link and semi-structured information. The organisers of INEX 2009 noticed that algorithms exhibiting high performance on Wikipedia were ineffective on a different Wiki collection [Huang et al.,2009] (mainly because it was not as extensively linked as Wikipedia and the title information was not as reliable). Similar findings have also been reported by [Erbs et al.,2011] who explored link discovery in corporate Wikis and found out that the information about the link graph was not helping the system as much as in the case of Wikipedia. As a result, we believe that link discovery evaluation workshops should in the future more encourage the development of methods that are applicable in a wider context. As these methods are unlikely to perform as well as methods specifically tailored for the collection used in the evaluation, there is currently little incentive to develop them and submit them for evaluation.

At the same time, the development of purely content-based approaches to CLLD constitute a number of challenges. In particular, (a) these approaches do not allow the use of cross-lingual links between Wikipedia articles - information that has been exploited and found very useful by most of the CrossLink participants, but can hardly be expected to be available in a general context, (b) anchor detection is a hugely challenging problem in a general context as links do not have to be of a conceptual nature (i.e. an anchor is connected to an article which explains it), but can constitute a whole range of semantic relationships [Knoth and Zdrahal,2011].

6. CONCLUSION

In this paper, we have presented and evaluated four different methods for Cross-Language Link Discovery (CLLD) applicable to the Wikipedia collection. We have used Explicit Semantic Analysis as a key component in the development of the four presented methods. Our methods produced good results as they ranked in all three evaluations in which we participated among the top three performers. The results suggest that methods that combine the knowledge of the Wikipedia link graph (including the cross-lingual mapping of articles) with textual semantic similarity can achieve promising results. However, such information is not generally applicable across textual collections and, therefore, it is reasonable to experiment with CLLD methods that operate at the level of textual content.

References

James Allan. 1997. Building hypertext using information retrieval. *Inf. Process. Manage.*, 33:145–159, March.

Philipp Dopichaj, Andre Skusa, and Andreas Heß. 2008. Stealing anchors to link the wiki. In Geva et al. [Geva et al.,2009], pages 343–353.

Nicolai Erbs, Torsten Zesch, and Iryna Gurevych. 2011. Link discovery: A comprehensive analysis. In *Proceedings of the 5th IEEE International Conference on Semantic Computing (IEEE-ICSC)*, page to appear, Jul.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors. 2009. *Advances in Focused Retrieval, 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008, Dagstuhl Castle, Germany, December 15-18, 2008. Revised and Selected Papers*, Lecture Notes in Computer Science. Springer.

Shlomo Geva. 2007. Gpx: Ad-hoc queries and automated link discovery in the wikipedia. In Norbert Fuhr, Jaap Kamps, Mounia Lalmas, and Andrew Trotman, editors, *INEX*, Lecture Notes in Computer Science. Springer.

Michael Granitzer, Christin Seifert, and Mario Zechner. 2008. Context based wikipedia linking. In Geva et al. [Geva et al.,2009], pages 354–365.

Stephen J. Green. 1998. Automated link generation: can we do better than term repetition? *Comput. Netw. ISDN Syst.*, 30(1-7):75–84.

Jiyin He. 2008. Link detection with wikipedia. In Geva et al. [Geva et al.,2009], pages 366–373.

Darren Wei Huang, Yue Xu, Andrew Trotman, and Shlomo Geva. 2008. Focused access to xml documents. chapter Overview of INEX 2007 Link the Wiki Track, pages 373–387. Springer-Verlag, Berlin, Heidelberg.

Wei Che Huang, Shlomo Geva, and Andrew Trotman. 2009. Overview of the inex 2009 link the wiki track.

Kelly Y. Itakura and Charles L. A. Clarke. 2008. University of waterloo at inex 2008: Adhoc, book, and link-the-wiki tracks. In Geva et al. [Geva et al.,2009], pages 132–139.

Dylan Jenkinson, Kai-Cheung Leung, and Andrew Trotman. 2008. Wikisearching and wikilinking. In Geva et al. [Geva et al.,2009], pages 374–388.

Petr Knoth and Zdenek Zdrahal. 2011. Mining cross-document relationships from text. In *The First International Conference on Advances in Information Mining and Management (IMMM 2011)*.

Petr Knoth, Jakub Novotny, and Zdenek Zdrahal. 2010. Automatic generation of inter-passage links based on semantic similarity. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 590–598, Beijing, China, August.

Petr Knoth, Lukas Zilka, and Zdenek Zdrahal. 2011a. Using explicit semantic analysis for cross-lingual link discovery. In *Workshop: 5th International Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies (CLIA) at The 5th International Joint Conference on Natural Language Processing (IJC-NLP 2011)*, Chiang Mai, Thailand, November.

Petr Knoth, Lukas Zilka, and Zdenek Zdrahal. 2011b. Using explicit semantic analysis for cross-lingual link discovery. In *International Workshop on Cross Lingual Information Access: : Computational Linguistics and the Information Need of Multilingual Societies (CLIA) at The 5th International Joint Conference on Natural Language Processing (IJC-NLP 2011)*, Chiang Mai, Thailand, November.

Wei Lu, Dan Liu, and Zhenzhen Fu. 2008. Csir at inex 2008 link-the-wiki track. In Geva et al. [Geva et al.,2009], pages 389–394.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA. ACM.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

Philipp Sorg and Philipp Cimiano. 2008. Enriching the crosslingual link structure of wikipedia - a classification-based approach -. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WikiAI'08)*, To appear.

Ling-Xiang Tang, Shlomo Geva, Andrew Trotman, Yue Xu, and Kelly Itakura. 2011. Overview of the ntcir-9 crosslink task: Cross-lingual link discovery. In *NTCIR-9*.

Andrew Trotman, David Alexander, and Shlomo Geva. 2009. Overview of the inex 2010 link the wiki track.

Jihong Zeng and Peter A. Bloniarz. 2004. From keywords to links: an automatic approach. *Information Technology: Coding and Computing, International Conference on*, 1:283.

Junte Zhang and Jaap Kamps. 2008. A content-based link detection approach using the vector space model. In Geva et al. [Geva et al.,2009], pages 395–400.